MAURI ORA, 1974, 2: 45-58

ORIGIN OF THE GENETIC CODE

D.C. REANNEY

Biochemistry Department, Lincoln College,
Canterbury, New Zealand

ABSTRACT

Data relating to the question of the origin of the genetic code are reviewed. It is considered likely that the primordial code(s) employed only two bases, I and U. A doublet code evolved in which coding specificity was confined to bases 1 and 2 of codons, the third base acting as a non-specific stabiliser. The primary interactions which first established a 'genetic' code probably employed a few amino acids only - those which could be accommodated by early transfer RNA-like polymers without the intervention of charging enzymes. However these early 'stereochemical' interactions need not necessarily be reflected in the structure of the contemporary code. The ordered structure of the modern code is considered to have arisen largely from stochastic processes.

INTRODUCTION

Even before the list of trinucleotide assignments had been completed certain regularities in the structure of the genetic code had generated a number of theories as to the code's origin (Jukes 1965, Woese 1965). Today, although the structure of the code has passed into the common body of scientific knowledge, the question of its origin remains a puzzle. Two general hypotheses have been advanced. The stochastic model was made possible by the adaptor hypothesis (Crick 1958), which seemed to separate amino acids and polynucleotides as 'complementary' entities. This model stresses the fitness of the contemporary code to buffer the phenotype of organisms against the effects of harmful mutation (Sonneborn 1965) and to increase the reliability of information transfer (Goldberg and Wittes 1966). The stereochemical model (see Woese 1967) implies that certain features of the code were uniquely specified by the physical nature of the first interacting components. Both theories have usually been presented as black and white alternatives. The fact that they are not mutually exclusive has seldom been explicitly stated (Reanney and Ralph 1967, Crick 1968).

THE GENETIC CODE CAN BE PRESENTED AS THE COMPLEMENT OF THE CODE ON mRNA

The genetic code can be expressed in three ways:

1. as the triplet of bases on DNA which specifies the complementary codon on mRNA,
2. as the codon on mRNA which interacts with the anti-codon, of tRNA, and
3. as the anticodon of tRNA.

All of these are formally valid representations of the molecular specificity of the code. The reasons why the code is usually written in terms of codons on mRNA are probably historical and may reflect the experimental techniques used to elucidate codons (for example the triplet binding assay used tRNAs of (largely) unknown primary structure; the known base sequences were those of the mRNA triplets). However in discussions of the code's origin it is more logical to represent the code in terms of the triplet of bases on tRNA, as the polynucleotides which interact with amino acids are tRNAs.

In Table 1 the code is presented in terms of the anticodons on tRNA written 3' → 5' to facilitate comparison with more usual coding catalogues. The bases in position 3 of anticodons are those known to occur in this position from determined primary structures of tRNAs and those predicted by the wobble hypothesis (Crick 1966).

TABLE 1. THE GENETIC CODE EXPRESSED IN TERMS OF THE CODING TRIPLETS ON tRNA

1st position	2nd position				3rd position	
	U	C	A	G	deaminated	aminated
U	ASN	SER	ILE	THR	I	
	ASN	SER	ILE	THR		G
	LYS	ARG	ILE	THR	U	
	LYS	ARG	MET	THR		C
C	ASP	GLY	VAL	ALA	I	
	ASP	GLY	VAL	ALA		G
	GLU	GLY	VAL	ALA	U	
	GLU	GLY	VAL	ALA		C
A	TYR	CYS	PHE	SER	I	
	TYR	CYS	PHE	SER		G
	TER	TER	LEU	SER	U	
	TER	TRP	LEU	SER		C
G	HIS	ARG	LEU	PRO	I	
	HIS	ARG	LEU	PRO		G
	GLN	ARG	LEU	PRO	U	
	GLN	ARG	LEU	PRO		C

Wobble pairing in base 3 (Crick 1966)

Codon	Anticodon
U C A	I
G	C
U C	G
A G	U

Note that according to this representation, inosine (I) is a letter in the alphabet of the code.

NO PHYSICAL EVIDENCE ON THE CODE'S ORIGIN IS LIKELY TO BE FOUND IN MOLECULAR PALEONTOLOGY

Recent developments in molecular paleontology make it possible to extend the direct study of trace remains of living things far back into the pre-Cambrian era. The oldest rocks investigated so far are the Fig Tree Chert series from South Africa (3100 million years) and the Onverwacht series, also in South Africa (3200 million years).

Engel *et al.* (1968) have found carbonaceous alga-like bodies in rocks of the Onverwacht series; they claim that these traces represent biological remains over 3.2 billion years old. These and other similar claims for ancient rocks have been treated with some scepticism by the scientific community because of the difficulties of unambiguously identifying any micro-fossil or chemical residue as 'biological' (see Calvin 1969). However, many, if not most workers in this field would agree with Engel *et al.* in their conclusions that "if the carbonaceous forms (in the Onverwacht series) are fossils, the origin of unicellular life presumably has occurred in still older rocks destroyed by superimposed igneous and metamorphic episodes in the evolving earth".

SOME EVIDENCE ON THE CODE'S ORIGIN MAY BE FROZEN IN THE BIOCHEMISTRY OF CONTEMPORARY ORGANISMS

In some cases enzymatic and other functions can be altered by mutation without a lethal effect upon the cell concerned. However if a given function has a large enough number of dependent reactions then any change in the basic function cannot be tolerated.

This gives rise to a concept of 'evolutionary conservatism' (Eck and Dayhoff 1966). Since the entire biochemistry of the cell depends upon the production, through the translation mechanism, of functional enzymes, it follows from the 'conservatism' principle that many basic features of the translation system cannot have changed throughout the measurable course of evolution. The code itself has probably been 'frozen' in a form rather similar to its present form for even longer, for similar reasons.

Thus it is not unreasonable to expect that some features of the most primitive amino acid-polynucleotide interactions survive as 'molecular fossils' in the most intimate biochemical reactions of cells and in the structure of the code itself.

THE NATURE OF EARLY CODES

ONLY A FEW AMINO ACIDS WERE LIKELY TO HAVE BEEN CODED IN ANCIENT CODES

While it is difficult to assess the significance of abiotic synthesis data, it does seem that certain amino acids are formed more readily than others. Of the amino acids

formed in typical experiments (Calvin 1969), the most common products from two independent reactions were GLY, ALA and the dicarboxylic acids, GLU and ASP. GLY, ALA, CYS, SER, ASP and VAL are perhaps the most thermodynamically stable of the amino acids (Eck and Dayhoff 1966). It is interesting that the 6 protein amino acids detected in the Murchison meteorite were GLY, ALA, GLU, ASP, VAL and PRO (Kvenvolden, Lawless and Ponnamperuma 1971). GLY, ALA, GLU and ASP are among the most abundant constituents of contemporary proteins.

It is widely agreed that TRP, MET and the amides of the dicarboxylic acids entered relatively late into the structure of proteins (Jukes 1965, Crick 1968).

IN INTERMEDIATE CODES 2 BASES ONLY MAY HAVE BEEN REQUIRED FOR AMINO ACID RECOGNITION

In the contemporary code, degeneracy is confined to the third letter. Jukes (1965) suggested that the archetypal code may have been a doublet code. The possibility that doublets alone can specify amino acids has been confirmed for SER by Rottman and Nirenberg (1966) who showed that the dinucleotide pUpC promoted the binding of SER tRNA to ribosomes in the triplet binding assay. There is now abundant evidence that individual tRNA species can recognise synonym codes differing only in the third letter and a theoretical basis for the observed pattern of degeneracy has been proposed by Crick (1966).

Careful examination of the data used to construct the present coding assignments reveals doublet regularities more extensive than those normally recognised (see below). These support the concept that in earlier forms of the genetic code the third base of the triplet was used solely as a stabilising base (Reanney and Ralph 1966). This being so, it is difficult to avoid the inference that less sophisticated codes used only two bases for specific recognition of their complements on mRNA.

Söll *et al.* (1965) showed that when challenged in the triplet binding assay with a variety of triplets, the 15-16 amino acids which are fairly widely accepted as 'primitive' (see Jukes 1965) gave 83 'significant' stimulations. Many triplets stimulated the binding of more than one amino acid. Reanney and Ralph (1966) were able to account for 81 of these *in vitro* binding stimulations on the basis of a doublet code.

The doublet code in Table 2 is derived from the coding assignments of Brimacombe *et al.* (1965) and Söll *et al.* (1965). In Table 2A the first two bases shown (the specificity doublet) are invariant while the third base (N) may vary.

Triplets in which the standard pattern of degeneracy in position 3 is preserved account for 54 of the 83 'significant' stimulations in Söll's assays. Certain triplets in which the specificity doublet occurs in position 1 and 2 do not fit the currently accepted pattern of degeneracy in place 3, e.g. UGA for CYS (standard codons UG-pyrimidine) and AAU for LYS (standard codons AA-purine). These triplets account for an additional 4 binding stimulations (Söll *et al.* 1965).

TABLE 2. INTERMEDIATE CODES REQUIRING TWO BASES ONLY FOR AMINO ACID RECOGNITION

A. THE DOUBLET CODE

SPECIFICITY DOUBLET	AMINO ACID	SPECIFICITY DOUBLET	AMINO ACID
UUN	PHe and LEU	AUN	ILE
UCN	SER*	ACN	THR*
UAN	TYR or Terminate	AAN	LYS
UGN	CYS	AGN	SER and ARG
CUN	LEU*	GUN	VAL*
CCN	PRO*	GCN	ALA*
CAN	HIS	GAN	GLU and ASP
CGN	ARG*	GGN	GLY*

* amino acids with 4 place degeneracy

AMINO ACIDS NOT CODED FOR IN A DOUBLET CODE[1]

AUG	MET
UGG	TRP
CA (A,G)	GLN
AA (A,C)	ASN

1 These amino acids are thought to have evolved later than those shown in the doublet code (see Jukes 1965).

B. SPECIFICITY DOUBLET: BASES 2 AND 3, DEGENERACY IN BASE 1

ALA	AGC		GLU	UGA		VAL	CGU		GLY	(CGG)	
	UGC	1		GGA			UGU			UGG	
	CGC	1		AGA	1	LYS	(AAA)			CGG	
ARG	CCG		HIS	UCA			UAA			AGG	
	GCG			CCA							
ASP	UGA		THR	CAC							
	GGA			AAC							

1 See Nirenberg *et al.* 1965.
All other assignments from Söll *et al.* 1965.

C. INVERTED SPECIFICITY DOUBLETS

ARG	GGC	CYS	CGU
	UGA	TYR	CAU
ASP	UAG		
	CAG		

All assignments are from Söll *et al.*

A further 17 binding stimulations in Söll's assays may be accounted for on the assumption that the specificity doublet occurs in positions 2 and 3, rather than 1 and 2 (Table 2B). While the position of the specificity doublet is altered in these triplets, the specificity of any doublet for its cognate amino acids remains unchanged from that in Table 2A.

Triplets with first or third letter degeneracy account for 75 of the triplets in Söll's table. This leaves a residue of 8 anomalous stimulations. However, 6 of these anomalous triplets contain specificity doublets in which the position of the bases is inverted (Table 2C). The stimulation of ASP binding by UAG and CAG (normal codons 5'GAU 3' or 5' GAC 3') might then become intelligible. This stimulation could arise if the conditions of the *in vitro* binding assay permit some tRNAs to pair with triplets in a parallel fashion instead of in the usual antiparallel manner. The fact that inverted doublets give relatively weak stimulation may imply that this type of binding is less favourable than binding with antiparallel polarity.

The trinucleotide binding assay is open to a number of criticisms. In particular, depending on the Mg^{++} concentration, one triplet can stimulate binding of several aminoacyl tRNAs; thus the test enhances ambiguity (Khorana 1965). It should also be emphasized that a triplet with 5' and 3' termini within the same codon represents an artificial situation the cell is unlikely to encounter. However, the fact that the dinucleotide pUpC would bind only SER tRNA (Rottman and Nirenberg 1966) (UC being the specificity doublet for SER) and not LEU or ILE whose codons contain UC in the second and third positions, suggests that the triplet binding test is adequate to cope with specific doublet recognition even though variation in the position and polarity of the doublet is permitted under assay conditions.

STILL EARLIER CODES MAY HAVE CONTAINED ONLY TWO BASES

The data presented in Table 2 suggest that intermediate codes selected among amino acid R groups by employing $4^2 = 16$ specificity doublets, the third base being physically present but playing no role in specificity. If one goes further back into the past, it is possible to conceive (Reanney and Ralph 1967, Crick 1968 *int. al.*) of very primitive nucleic acid molecules which contained two bases only ($2^2 = 4$ specificity doublets).

It should be stressed that there is no experimental support for the claim of a two letter code; it is simply a plausible assumption which makes the question of the code's origin easier to understand (Crick 1968).

If one accepts the notion of a two letter code, an important question is which two bases may have been involved. Some inferences can be drawn from abiotic synthesis; it does seem true for example that adenine is rather more readily formed than any other base. The resonance energy of adenine is greater than that for any other biologically important heterocyclic base (Pullman and Pullman 1962). It may also be true that, of the pyrimidines, uracil is more easily formed than cytosine or thymine (see Pattee 1965).

Crick (1968) has suggested that the primordial nucleic acid contained A and I, as a stable double helix can be formed from poly A and I and the more primitive amino acids do tend to have purines only in their specificity doublets (I codes like G but with one less H bond). I can be formed readily in abiotic synthesis (see Lowe, Rees and Markham 1963) but Crick suggests that it was perhaps formed from adenine by

deamination.

THE 2 INITIAL BASES MAY HAVE BEEN U AND I

This leads to a point which is seldom considered in discussions of the code's origin, i.e. that the unphysiological conditions needed to generate many monomers in abiotic experiments would also lead to their modification or breakdown over long periods. It is known from the chemistry of the nucleic acids that the bases are prone to deamination in a number of circumstances. Under non-oxidising conditions at acid pH, C is converted into U (Hunter and Hlynka 1937). Alkaline treatment of the purine nucleosides at high temperatures for one hour leads to 8-9% deamination (Jones, Mian and Walker 1966).

Reanney and Ralph (1967) have postulated that the two (predominant) original bases in polynucleotide were the deaminated bases U and I. Several advantages follow from this hypothesis:

1. these two bases may have been more stable (above), hence more readily available as precursors under primitive conditions than the aminated bases;
2. in contemporary biosynthetic pathways U and I precede the other nucleotides (Fig. 1); their manufacture would thus be 'cheaper' in terms of energy and enzymes (the fact that the amination which leads to C, A and G occurs after the formation of UMP and IMP might imply that this amination step arose later in the evolution of this biosynthetic pathway).
3. in certain *in vitro* situations polymers containing I will replicate whereas those containing G will not (Karstadt and Krakow 1970), perhaps because the Tm of duplex molecules containing G/C pairs prohibits effective strand separation (Beibricher and Orgel 1973).
4. a double helix of the standard Watson-Crick type can in theory be constructed from poly I and poly U or IU copolymers by invoking enol tautomerism of I (Reanney and Ralph 1967).

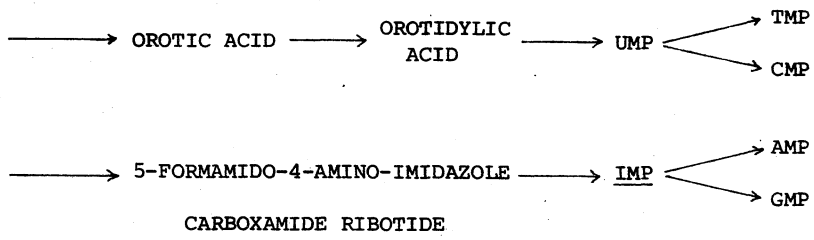

Fig. 1. Biosynthetic pathways leading to the mononucleotides.

The poly AI helix postulated by Crick suffers from the disadvantage that both A and I are purines. Orgel (1968) has pointed out that the only polynucleotides of possible evolutionary relevance are those whose complements also serve an efficient template function. In aqueous solution poly U cannot be formed on a poly A template because of the poor stacking ability of the pyrimidines (Sulston *et al.* 1968). The introduction of the pyrimidines into a predominantly A-I

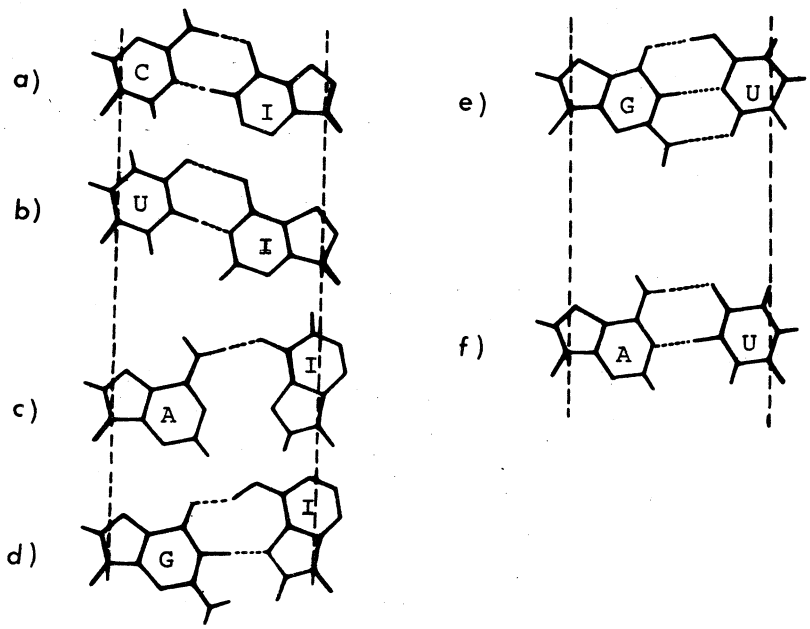

Fig. 2. Alternative base pairs which preserve the standard interglycosidic bond distance of DNA.

Footnote to Fig. 2: DOUBLE HELICES POSSIBLE WITH U AND I WHICH PRESERVE THE STANDARD INTERGLYCOSIDIC BOND DISTANCE OF DNA. ADAPTED FROM REANNEY AND RALPH (1967).

Figures 2a, b, c and d show how the 4 standard nucleotides might pair with I. The pK of the 6 substituent of inosine (pK 8.75) is such that it is partly ionised at pH 7.8. The pK may be even lower than when I in a polynucleotide is base-paired to other mononucleotides. I can make only a single bond with A as shown in 2c without distorting the distance between the glycosidic bonds in the base pair. This configuration (2c) might permit I in a polynucleotide to bond to A in a second polynucleotide if the bases adjacent to I were base paired. However I in a polynucleotide would probably not form a standard double helix with A in this configuration. I paired to A in any other way excessively distorts the glycosidic bond distances. The tautomeric form of I may require the concerted approach of the appropriately oriented H bonding groups of U to facilitate and maintain this form of bond. Thus it is unlikely that the tautomeric form of I would be elicited to bond to G as shown in 2d. U is able to form 3 H bonds with G when the enol form of U is bonded as shown in 2e. This does not unnecessarily distort the interglycosidic bond distances and the resulting tautomeric form of U might be facilitated and stabilised by the concerted action of the 2-amino and 6-keto groups of G to which it becomes base paired. Alternatively, as suggested for I, stabilisation of the U-G pair might result from the existence of some ionised U at physiological pH.

The arguments put forward here gain substance when it is remembered that the presence of organic molecules in the environment of primitive polynucleotides may have enhanced enolisation (So and Davie 1964).

helix would interrupt the geometry of the helix as the inter-glycosidic bond distance A-I is greater than the distances A-T(U) or I-C. By contrast, C, A and G could be introduced into a poly IU helix without any distortion of the molecule (Fig. 2). However, Arnott and Bond (1973) have shown that in a triple stranded polynucleotide helix involving one poly A and two poly I chains, the three helical chains have conformations similar to conventional A-type double helices despite the absence of the standard purine-pyrimidine base pairs. They suggested that the evolution of the contemporary genetic code from a primitive code in which A and I were the predominant bases would not have required 'major discontinuities' in molecular geometry (Arnott and Bond 1973).

A major supporting prop for the notion that U and I were the original bases is the inference (Reanney and Ralph 1967) that, in cyanimide type condensations, only the deaminated bases (especially in combination with deoxy sugars) could have formed linear, unbranched 3' → 5' linked polynucleotides. For a fuller discussion of the chemistry of this point the reader is referred to Reanney and Ralph (1967), Tener *et al.* (1958) and Ralph and Khorana (1961).

A POLY UI CODE CAN EXPLAIN THE ORIGIN OF THE PATTERN OF DEGENERACY IN BASE 3

Fig. 3 represents a summary of the events in the postulated evolution of polynucleotides based on U and I to those employing the 4 standard bases. Presumably the animated bases would have been introduced into adaptor polynucleotide over the same time period as into template.

1. Formation of		pI + pU
2. Formation of template (1)	(1)	pI pU pI pU pU pI pU pI pU pU
3. Formation of		pA + pC + pG
4. Formation of template (2) by condensation of pU + pA + pC + pG on template (1)	(1)	pI pU pI pU pU pI pU pI pU pU
	(2)	pU pA pC pA pG pU pA pC pA pG
5. Formation of template (3) on template (2)	(2)	pU pA pC pA pG pU pA pC pA pG
	(3)	pA pU pG pU pC pA pU pG pU pC

Fig. 3. Summary of events during the formation, replication and evolution of polynucleotide templates.

This evolution can explain in an unforced way the origin of the present pattern of degeneracy. To understand this, one must look briefly at the interaction of mRNA and tRNA on the ribosome. As the ribosomal site(s) for tRNA must accommodate all tRNA species, each individual tRNA must bind in a strictly equivalent way; specificity is confined exclusively (?) to the H bonding interaction between codon and anticodon. To serve as a basis for specificity and yet allow rapid flux of components through the ribosome the

interaction must be strong enough to hold the tRNA to the mRNA-ribosome complex but weak enough not to hinder easy release. The weakest interaction possible in contemporary systems is the bonding UUU (codon) - AAG (anticodon). This has 4 standard H bonds and one weaker H bond. The strongest interaction would seem to be GGG - CCC (9 standard H bonds). It has yet to be proven that this latter interaction actually takes place *in vivo*.

The permitted range seems to be the bond energy of 4-5 → 8-9 bonds. It is a reasonable assumption that the function of the third 'spacer' base is/was less to provide specificity (although it does this to some extent in the contemporary code) than to provide the extra energy necessary to stabilise the codon-anticodon interaction in the correct Tm range.

Fig. 2 shows that U could pair ambiguously with A and G while I could pair ambiguously with C or U(T). As the aminated bases became available, they would rapidly replace U and I in the specificity doublets of codons, owing to the possibilities for greater specificity and the introduction into the code of new amino acids; however such pressures would not apply to the third base. Retention of ambiguous pairing here would lead to a situation where the two purines and two pyrimidines were not discriminated in base 3 of all sense codons.

As new amino acids spread over the code some amino acids whose role in protein structure was more crucial than most (e.g., PRO) or whose frequency in 'statistical' protein was high (e.g., GLY) would acquire genetic 'protection' against mutational alteration in the form of extra codons. This follows from the predictions of all stochastic models of the code's evolution (Goldberg and Wittes 1966, Crick 1968). Four place degeneracy, i.e. four codons rather than two, could be achieved using U and I only; this is still a possibility in the contemporary code (Table 1). The retention of I into the modern code or its re-introduction by eukaryotes following the development of 'anticodon deaminase' follows from the wobble pairing possible with I (Crick 1966). C was probably introduced into base 3 of anticodons only when the 'late' amino acids MET and TRP entered the code. MET in particular has come to play a unique role in the initiation of protein biosynthesis; its assignment of a unique codon (AUG) would remove this triplet from the set presumably previously occupied by ILE. The unusual pattern of degeneracy shown by ILE would follow from this hypothesis. It is interesting that the two late additions to the code are both concerned with punctuation; MET with initiation, TRP taking over a codon which was probably previously a TER codon (UGG obeys the rule: U.pu.pu in mRNA = TER). This suggests that there occurred at some stage of evolution a significant modification of the system of translation. The observation that certain viral proteins are derived by enzymatic hydrolysis from a much larger polypeptide suggests that primitive chromosomes were translated as wholes and that the active 'fragments' arose from hydrolysis at 'weak links' in the chain. In 28 out of 29 cytochromes, the C terminal amino acids are related to chain terminating codons by a one base change (Nolan and Margoliash 1968).

It may be significant that the 8 amino acids with 4 place degeneracy all contain G and/or C in their specificity doublets, suggesting that extra stabilisation was required to compensate for the proposed nonstandard pairings in position 3 of triplets.

The replacement of U by T would follow from the fact that the higher pK of T, (pK 9.8) cf. U (pK 9.2), reduced the possibility of ambiguous pairing (Reanney and Ralph 1967).

PROTO tRNA

The pre-requisites for any primitive adaptor are:

1. a terminus (not necessarily CCA) to which the amino acid can be transferred, and
2. a region of complementarity (not necessarily involving the current bases) which can base pair with message.

It also seems likely that such a molecule would probably require some degree of 3D structure. Thus one is forced to the (at first sight) unlikely conclusion that even the most primitive adaptors were tolerably complicated molecules. On the basis of sequence homologies Jukes (1966) has postulated that all known tRNAs derive from a common ancestor 84 nucleotides long. It is hard to conceive of the most primitive translation systems using so elaborate a molecule. Even if sequence periodicity within present tRNAs indicates, however, that the 'ancestral' tRNA was derived by gene duplication from a smaller gene (or gene set), the requirement for tertiary structure sets a limit to the 'smallness' of the adaptor.

THE DIFFICULTY OF DISPROVING THE STEREOCHEMICAL HYPOTHESIS

One can now invert current arguments concerning the likelihood of discovering amino acid-polynucleotide interactions. It would in fact have been surprising if any such interactions had been discovered, given the reaction conditions and components hitherto employed.

It is important to note that the stereochemical model does not require polynucleotides to recognise all 20 amino acids. In contrast with some protagonists of the stereochemical theory I believe that stereochemical interactions were important only for the few amino acids necessary to establish the system. This presupposes that most attempts to demonstrate some kind of interaction are doomed to failure at the outset. Furthermore it is by no means certain that the polymers implicated in the initial interactions resembled modern nucleic acids in terms of component bases or even in terms of the 3'–5' phosphodiester linkage (Sulston *et al.* 1968). The point of these considerations is to stress that the later evolution of the code may have obliterated many of the features on which its inception depended (but see 4).

Fresco *et al.* (1966) have convincingly demonstrated that the biological activity of tRNA can be critically affected by factors such as pH, temperature, electrolyte concentration and Mg levels. The chemical milieu in which the initial coding interactions took place was undoubtedly very different to the reaction mixtures used by modern experimentalists. The presence of a variety of organic molecules (formaldehyde,

ethanol, etc.) may have conferred upon the solution solvent properties lacking in 'physiological' media. For example organic compounds such as ethanol might enhance the likelihood of tautomerism in the bases (So and Davie 1964).

Thus unless it is possible to specify more exactly the nature of the environment of the interacting components, then it may be unrealistic to expect any clear-cut binding specificity between amino acids (and/or their derivatives) and polynucleotides to reveal itself.

STOCHASTIC MODELS OF THE CODE'S EVOLUTION

A number of workers have stressed that the contemporary code is highly nonrandom. Mackay (1967) showed that the natural code closely approximated to a statistically constructed 'optimal' code. Alff-Steinberger (1969) compared the amino acid substitutions resulting from single base substitutions in the natural code with those of substitutions in computer generated random codes. For a number of amino acid parameters (molecular weight, polar requirement, pKi, etc.) it was shown that single base substitutions in the first position of codons tended to result in the substitution of an amino acid more similar to the original amino acid than would be expected from a random code. The 'protective' features of the code have been convincingly summarised by Goldberg and Wittes (1966).

Thus the code can be considered (in the jargon of cybernetics) as an error-minimising code (Alff-Steinberger 1969). It follows that the code may have been brought into being solely by the constant selective pressures to ensure fidelity of translation and to buffer the phenotype of organisms against the effects of mutation.

Features of stochastic models have been reviewed by Woese (1967). I will confine myself to one aspect of the code which (to the best of my knowledge) has not been treated so far in stochastic arguments on code evolution.

In the modern code degeneracy is confined to base 3 but an equally 'protective' code is theoretically possible with the redundancy shifted to base 1. The localisation of degeneracy in base 3 may arise from the vectorial nature of translation: if the probability of mutation is randomly distributed along the length of the genome, then it is advantageous to have the two specific bases in which the identity of the amino acid resides as close as possible to the N terminus from which peptide synthesis starts. This arrangement maximises the chance that a given base sequence will be 'safely' translated before mutation can change the identity of the amino acid(s).

It is very likely that much of the structure of the contemporary code results from selective pressure as predicted by stochastic models. But while stochastic arguments can convincingly explain the later evolution of the code, they say nothing about its origin. It is not easy to see how any system of translation could have begun from the completely random assortment of amino acid - nucleotide interactions which the stochastic argument, in its extreme form, implies. It is to be hoped that some stereochemical specificities can be experimentally demonstrated otherwise our chances of understanding the code's origin are remote.

LITERATURE CITED

ALFF-STEINBERGER, C. 1969. The genetic code and error transmission. *Proceedings of the National Academy of Sciences (USA) 64*: 584-591.

ARNOTT, S. and BOND, P.J. 1973. Triple stranded polynucleotide helix containing only purine bases. *Science 181*: 68-69.

BIEBRICHER, C.K. and ORGEL, L.E. 1973. An RNA that multiplies indefinitely with DNA dependent RNA polymerase: selection from a random copolymer. *Proceedings of the National Academy of Sciences (USA) 70*: 934-938.

BRIMACOMBE, R., TRUPIN, J., NIRENBERG, M., LEDER, P., BERNFIELD, M. and JAOUNI, T. 1965. RNA codewords and protein synthesis. VIII. Nucleotide sequences of synonym codons for arginine, valine, cysteine and alanine. *Proceedings of the National Academy of Sciences (USA) 54*: 954-960.

CALVIN, M. 1969. *Chemical evolution.* Oxford (Clarendon Press), London. 278 pp.

CRICK, F.H.C. 1966. Codon-anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology 19*: 538-555.

———— 1968. The origin of the genetic code. *Journal of Molecular Biology 38*: 367-379.

ECK, R. and DAYHOFF, M. 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science 152*: 363-366.

ENGEL, A.E.J., NAGY, B., NAGY, L.A., ENGEL, C.G., KREMP, A.W.W., and DREW, C.M. 1968. Alga-like forms in Onverwacht series, South Africa: oldest recognised lifelike forms on Earth. *Science 161*: 1005-1008.

FRESCO, J.R., ADAMS, A., ASCIONE, R., HENLEY, D. and LINDAHL, T. 1966. Tertiary structure in transfer ribonucleic acids. *Cold Spring Harbor Symposia on Quantitative Biology. XXXI. The genetic code*: 527-537.

GOLDBERG, A.L. and WITTES, R.E. 1966. Genetic code: aspects of organisation. *Science 153*: 420-424.

HUNTER, A. and HLYNKA, I. 1937. Note on the preparation of purines and pyrimidines from nucleic acid. *Biochemical Journal 31*: 486-487.

JONES, A.S, MIAN, A.M. and WALKER, R.I. 1966. The action of alkali on some purines and their derivatives. *Journal of Chemical Science C*: 692-695.

JUKES, T.H. 1965. Coding triplets and their possible evolutionary implications. *Biochemical and Biophysical Research Communications 19*: 391-396.

———— 1966. Indications for a common evolutionary origin shown in the primary structure of three transfers RNA's. *Biochemical and Biophysical Research Communications 24*: 744-749.

KARSTADT, M. and KRAKOW, J.S. 1970. *Azotobacter vinelandii* ribonucleic acid polymerase. *Journal of Biological Chemistry 245*: 746-751.

KHORANA, H.G. 1965. Polynucleotide synthesis and the genetic code. *Federation Proceedings 24*: 1473-1487.

KVENVOLDEN, K., LAWLESS, J.C. and PONNAMPERUMA, C. 1971. Nonprotein amino acids in the Murchison meteorite. *Proceedings of the National Academy of Science (USA) 68*: 486-490.

LOWE, C.U., REES, M.W. and MARKHAM, R. 1963. Synthesis of complex organic compounds from simple precursors: Formation of amino acids, amino acid polymers, fatty acids and purines from ammonium cyanide. *Nature 199*: 219-222.

MACKAY, A.L. 1967. Optimization of the genetic code. *Nature 216*: 159-160.

ORGEL, L.E. 1968. Evolution of the genetic apparatus. *Journal of Molecular Biology 38*: 381-393.

NIRENBERG, M., LEDER, P., BERNFIELD, M., BRIMACOMBE, R., TRUPIN, J., ROTTMAN, F. and O'NEAL, C. 1965. RNA codewords and protein synthesis. VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences (USA) 53*: 1161-1168.

NOLAN, C. and MARGOLIASH, E. 1968. Primary structures of proteins. *Annual Reviews of Biochemistry 37*: 727-790.

PATTEE, H.H. 1965. Experimental approaches to the origin of life problem. *Advances in Enzymology 27*: 381-415.

PULLMAN, B. and PULLMAN, A. 1962. Electronic delocalization and biochemical evolution. *Nature 196*: 1137-1142.

RALPH, R.K. and KHORANA, H.G. 1961. Studies on polynucleotides. XI. Chemical polymerisation of mononucleotides. The synthesis characterisation of deoxyadenosine polynucleotides. *American Chemical Society Journal 83*: 2926-2934.

REANNEY, D.C. and RALPH, R.K. 1966. The possible evolutionary significance of doublet regularities in the code. *National Institute of Health Information Exchange Group. No. 7 - memo no. 436.*

———————————— 1967. A speculation of the origin of the genetic code. *Journal of Theoretical Biology 15*: 41-52.

ROTTMAN, F. and NIRENBERG, M. 1966. RNA codons and protein synthesis. XI. Template activity of modified RNA codons. *Journal of Molecular Biology 21*: 555-570.

SO, A.G. and DAVIE, E.W. 1964. The effects of organic solvents on protein biosynthesis and their influence on the amino acid code. *Biochemistry 3*: 1165-1169.

SÖLL, D., OHTSUKA, E., JONES, D., LOHRMANN, R., HAYATSU, H., NISHIMURA, S. and KHORANA, H. 1965. Studies of polynucleotides. XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribonucleotides and a survey of codon assignments for 20 amino acids. *Proceedings of the National Academy of Sciences (USA) 54*: 1378-1385.

SONNEBORN, T.M. 1965. Degeneracy of the genetic code: extent, nature and genetic implications. *In*: Bryson, V. and Vogel, H.J. (Eds), *Evolving genes and proteins*: 377-397. Academic Press, New York.

SULSTON, J., LOHRMANN, R., ORGEL, L.E. and MILES, H.T. 1968. Specificity of oligonucleotide synthesis directed by polyuridylic acid. *Proceedings of the National Academy of Sciences (USA) 60*: 409-415.

TENER, M.G., KHORANA, H.G., MARKHAM, R. and POL, E.H. 1958. Studies on polynucleotides. II. The synthesis and characterisation of linear and cyclic thymidine oligonucleotides. *American Chemical Society Journal 80*: 6223-6230.

WOESE, C.R. 1965. On the evolution of the genetic code. *Proceedings of the National Academy of Sciences (USA) 54*: 1546-1552.

———————— 1967. *The genetic code.* Harper and Rowe, New York.